Machine Learning Excercise

SP

29 december 2018

Introduction

In this exercise a dataset is studied related to personal activities measured by accelerometers at several bodyparts. The goal is to find a model that could predict whether or not the exercise was performed in the right way

Load and preprocessing of the data

At first the data should be retrieved and pre-processed.

```
pml.training <- read.csv(url("https://urldefense.proofpoint.com/v2/url?u=https-3A__d396qusza40orc.cloudf
pml.testing <- read.csv(url("https://urldefense.proofpoint.com/v2/url?u=https-3A__d396qusza40orc.cloudf
```

Reviewing the data shows that the pml.training set and pml.test set has both 160 variables and respectively 19622 and 20 observations. Analyzing the data in more detail show that there are a couple of columns in the dataset that don't have any relations with the accelerometers.For instance, 'x', 'user_name' and 'raw_timestamp_part_1'.

```
colnames(pml.training)
```

```
##   [1] "X"                        "user_name"
##   [3] "raw_timestamp_part_1"     "raw_timestamp_part_2"
##   [5] "cvtd_timestamp"           "new_window"
##   [7] "num_window"               "roll_belt"
##   [9] "pitch_belt"               "yaw_belt"
##  [11] "total_accel_belt"         "kurtosis_roll_belt"
##  [13] "kurtosis_picth_belt"      "kurtosis_yaw_belt"
##  [15] "skewness_roll_belt"       "skewness_roll_belt.1"
##  [17] "skewness_yaw_belt"        "max_roll_belt"
##  [19] "max_picth_belt"           "max_yaw_belt"
##  [21] "min_roll_belt"            "min_pitch_belt"
##  [23] "min_yaw_belt"             "amplitude_roll_belt"
##  [25] "amplitude_pitch_belt"     "amplitude_yaw_belt"
##  [27] "var_total_accel_belt"     "avg_roll_belt"
##  [29] "stddev_roll_belt"         "var_roll_belt"
##  [31] "avg_pitch_belt"           "stddev_pitch_belt"
##  [33] "var_pitch_belt"           "avg_yaw_belt"
##  [35] "stddev_yaw_belt"          "var_yaw_belt"
##  [37] "gyros_belt_x"             "gyros_belt_y"
##  [39] "gyros_belt_z"             "accel_belt_x"
##  [41] "accel_belt_y"             "accel_belt_z"
##  [43] "magnet_belt_x"            "magnet_belt_y"
##  [45] "magnet_belt_z"            "roll_arm"
##  [47] "pitch_arm"                "yaw_arm"
##  [49] "total_accel_arm"          "var_accel_arm"
```

```
##  [51] "avg_roll_arm"              "stddev_roll_arm"
##  [53] "var_roll_arm"              "avg_pitch_arm"
##  [55] "stddev_pitch_arm"          "var_pitch_arm"
##  [57] "avg_yaw_arm"               "stddev_yaw_arm"
##  [59] "var_yaw_arm"               "gyros_arm_x"
##  [61] "gyros_arm_y"               "gyros_arm_z"
##  [63] "accel_arm_x"               "accel_arm_y"
##  [65] "accel_arm_z"               "magnet_arm_x"
##  [67] "magnet_arm_y"              "magnet_arm_z"
##  [69] "kurtosis_roll_arm"         "kurtosis_picth_arm"
##  [71] "kurtosis_yaw_arm"          "skewness_roll_arm"
##  [73] "skewness_pitch_arm"        "skewness_yaw_arm"
##  [75] "max_roll_arm"              "max_picth_arm"
##  [77] "max_yaw_arm"               "min_roll_arm"
##  [79] "min_pitch_arm"             "min_yaw_arm"
##  [81] "amplitude_roll_arm"        "amplitude_pitch_arm"
##  [83] "amplitude_yaw_arm"         "roll_dumbbell"
##  [85] "pitch_dumbbell"            "yaw_dumbbell"
##  [87] "kurtosis_roll_dumbbell"    "kurtosis_picth_dumbbell"
##  [89] "kurtosis_yaw_dumbbell"     "skewness_roll_dumbbell"
##  [91] "skewness_pitch_dumbbell"   "skewness_yaw_dumbbell"
##  [93] "max_roll_dumbbell"         "max_picth_dumbbell"
##  [95] "max_yaw_dumbbell"          "min_roll_dumbbell"
##  [97] "min_pitch_dumbbell"        "min_yaw_dumbbell"
##  [99] "amplitude_roll_dumbbell"   "amplitude_pitch_dumbbell"
## [101] "amplitude_yaw_dumbbell"    "total_accel_dumbbell"
## [103] "var_accel_dumbbell"        "avg_roll_dumbbell"
## [105] "stddev_roll_dumbbell"      "var_roll_dumbbell"
## [107] "avg_pitch_dumbbell"        "stddev_pitch_dumbbell"
## [109] "var_pitch_dumbbell"        "avg_yaw_dumbbell"
## [111] "stddev_yaw_dumbbell"       "var_yaw_dumbbell"
## [113] "gyros_dumbbell_x"          "gyros_dumbbell_y"
## [115] "gyros_dumbbell_z"          "accel_dumbbell_x"
## [117] "accel_dumbbell_y"          "accel_dumbbell_z"
## [119] "magnet_dumbbell_x"         "magnet_dumbbell_y"
## [121] "magnet_dumbbell_z"         "roll_forearm"
## [123] "pitch_forearm"             "yaw_forearm"
## [125] "kurtosis_roll_forearm"     "kurtosis_picth_forearm"
## [127] "kurtosis_yaw_forearm"      "skewness_roll_forearm"
## [129] "skewness_pitch_forearm"    "skewness_yaw_forearm"
## [131] "max_roll_forearm"          "max_picth_forearm"
## [133] "max_yaw_forearm"           "min_roll_forearm"
## [135] "min_pitch_forearm"         "min_yaw_forearm"
## [137] "amplitude_roll_forearm"    "amplitude_pitch_forearm"
## [139] "amplitude_yaw_forearm"     "total_accel_forearm"
## [141] "var_accel_forearm"         "avg_roll_forearm"
## [143] "stddev_roll_forearm"       "var_roll_forearm"
## [145] "avg_pitch_forearm"         "stddev_pitch_forearm"
## [147] "var_pitch_forearm"         "avg_yaw_forearm"
## [149] "stddev_yaw_forearm"        "var_yaw_forearm"
## [151] "gyros_forearm_x"           "gyros_forearm_y"
## [153] "gyros_forearm_z"           "accel_forearm_x"
## [155] "accel_forearm_y"           "accel_forearm_z"
## [157] "magnet_forearm_x"          "magnet_forearm_y"
```

```
## [159] "magnet_forearm_z"         "classe"
```

Since the code above show the first 7 columns are not related to the goal of predicting, they are removed.

```
pml.testing <- pml.testing[,8:160]
pml.training <- pml.training[,8:160]
```

Scrolling through the dataset reveals also a lot of columns that have plenty of NA's in it.

```
NA_training <- sapply(pml.training, function(x) sum(is.na(x)))
NA_testing <- sapply(pml.testing, function(x) sum(is.na(x)))
col_nummer <- seq(1:153)
NA_df <- data.frame(as.data.frame(NA_training), as.data.frame(NA_testing), as.data.frame(col_nummer))
print(NA_df)
```

```
##                          NA_training NA_testing col_nummer
## roll_belt                          0          0          1
## pitch_belt                         0          0          2
## yaw_belt                           0          0          3
## total_accel_belt                   0          0          4
## kurtosis_roll_belt                 0         20          5
## kurtosis_picth_belt                0         20          6
## kurtosis_yaw_belt                  0         20          7
## skewness_roll_belt                 0         20          8
## skewness_roll_belt.1               0         20          9
## skewness_yaw_belt                  0         20         10
## max_roll_belt                  19216         20         11
## max_picth_belt                 19216         20         12
## max_yaw_belt                       0         20         13
## min_roll_belt                  19216         20         14
## min_pitch_belt                 19216         20         15
## min_yaw_belt                       0         20         16
## amplitude_roll_belt            19216         20         17
## amplitude_pitch_belt           19216         20         18
## amplitude_yaw_belt                 0         20         19
## var_total_accel_belt           19216         20         20
## avg_roll_belt                  19216         20         21
## stddev_roll_belt               19216         20         22
## var_roll_belt                  19216         20         23
## avg_pitch_belt                 19216         20         24
## stddev_pitch_belt              19216         20         25
## var_pitch_belt                 19216         20         26
## avg_yaw_belt                   19216         20         27
## stddev_yaw_belt                19216         20         28
## var_yaw_belt                   19216         20         29
## gyros_belt_x                       0          0         30
## gyros_belt_y                       0          0         31
## gyros_belt_z                       0          0         32
## accel_belt_x                       0          0         33
## accel_belt_y                       0          0         34
## accel_belt_z                       0          0         35
## magnet_belt_x                      0          0         36
## magnet_belt_y                      0          0         37
```

```
## magnet_belt_z                   0          0          38
## roll_arm                        0          0          39
## pitch_arm                       0          0          40
## yaw_arm                         0          0          41
## total_accel_arm                 0          0          42
## var_accel_arm               19216         20          43
## avg_roll_arm                19216         20          44
## stddev_roll_arm             19216         20          45
## var_roll_arm                19216         20          46
## avg_pitch_arm               19216         20          47
## stddev_pitch_arm            19216         20          48
## var_pitch_arm               19216         20          49
## avg_yaw_arm                 19216         20          50
## stddev_yaw_arm              19216         20          51
## var_yaw_arm                 19216         20          52
## gyros_arm_x                     0          0          53
## gyros_arm_y                     0          0          54
## gyros_arm_z                     0          0          55
## accel_arm_x                     0          0          56
## accel_arm_y                     0          0          57
## accel_arm_z                     0          0          58
## magnet_arm_x                    0          0          59
## magnet_arm_y                    0          0          60
## magnet_arm_z                    0          0          61
## kurtosis_roll_arm               0         20          62
## kurtosis_picth_arm              0         20          63
## kurtosis_yaw_arm                0         20          64
## skewness_roll_arm               0         20          65
## skewness_pitch_arm              0         20          66
## skewness_yaw_arm                0         20          67
## max_roll_arm                19216         20          68
## max_picth_arm               19216         20          69
## max_yaw_arm                 19216         20          70
## min_roll_arm                19216         20          71
## min_pitch_arm               19216         20          72
## min_yaw_arm                 19216         20          73
## amplitude_roll_arm          19216         20          74
## amplitude_pitch_arm         19216         20          75
## amplitude_yaw_arm           19216         20          76
## roll_dumbbell                   0          0          77
## pitch_dumbbell                  0          0          78
## yaw_dumbbell                    0          0          79
## kurtosis_roll_dumbbell          0         20          80
## kurtosis_picth_dumbbell         0         20          81
## kurtosis_yaw_dumbbell           0         20          82
## skewness_roll_dumbbell          0         20          83
## skewness_pitch_dumbbell         0         20          84
## skewness_yaw_dumbbell           0         20          85
## max_roll_dumbbell           19216         20          86
## max_picth_dumbbell          19216         20          87
## max_yaw_dumbbell                0         20          88
## min_roll_dumbbell           19216         20          89
## min_pitch_dumbbell          19216         20          90
## min_yaw_dumbbell                0         20          91
```

```
## amplitude_roll_dumbbell        19216          20           92
## amplitude_pitch_dumbbell       19216          20           93
## amplitude_yaw_dumbbell             0          20           94
## total_accel_dumbbell               0           0           95
## var_accel_dumbbell             19216          20           96
## avg_roll_dumbbell              19216          20           97
## stddev_roll_dumbbell           19216          20           98
## var_roll_dumbbell              19216          20           99
## avg_pitch_dumbbell             19216          20          100
## stddev_pitch_dumbbell          19216          20          101
## var_pitch_dumbbell             19216          20          102
## avg_yaw_dumbbell               19216          20          103
## stddev_yaw_dumbbell            19216          20          104
## var_yaw_dumbbell               19216          20          105
## gyros_dumbbell_x                   0           0          106
## gyros_dumbbell_y                   0           0          107
## gyros_dumbbell_z                   0           0          108
## accel_dumbbell_x                   0           0          109
## accel_dumbbell_y                   0           0          110
## accel_dumbbell_z                   0           0          111
## magnet_dumbbell_x                  0           0          112
## magnet_dumbbell_y                  0           0          113
## magnet_dumbbell_z                  0           0          114
## roll_forearm                       0           0          115
## pitch_forearm                      0           0          116
## yaw_forearm                        0           0          117
## kurtosis_roll_forearm              0          20          118
## kurtosis_picth_forearm             0          20          119
## kurtosis_yaw_forearm               0          20          120
## skewness_roll_forearm              0          20          121
## skewness_pitch_forearm             0          20          122
## skewness_yaw_forearm               0          20          123
## max_roll_forearm               19216          20          124
## max_picth_forearm              19216          20          125
## max_yaw_forearm                    0          20          126
## min_roll_forearm               19216          20          127
## min_pitch_forearm              19216          20          128
## min_yaw_forearm                    0          20          129
## amplitude_roll_forearm         19216          20          130
## amplitude_pitch_forearm        19216          20          131
## amplitude_yaw_forearm              0          20          132
## total_accel_forearm                0           0          133
## var_accel_forearm              19216          20          134
## avg_roll_forearm               19216          20          135
## stddev_roll_forearm            19216          20          136
## var_roll_forearm               19216          20          137
## avg_pitch_forearm              19216          20          138
## stddev_pitch_forearm           19216          20          139
## var_pitch_forearm              19216          20          140
## avg_yaw_forearm                19216          20          141
## stddev_yaw_forearm             19216          20          142
## var_yaw_forearm                19216          20          143
## gyros_forearm_x                    0           0          144
## gyros_forearm_y                    0           0          145
```

```
## gyros_forearm_z                     0          0         146
## accel_forearm_x                     0          0         147
## accel_forearm_y                     0          0         148
## accel_forearm_z                     0          0         149
## magnet_forearm_x                    0          0         150
## magnet_forearm_y                    0          0         151
## magnet_forearm_z                    0          0         152
## classe                              0          0         153
```

Since 19216 of the 19622 observations in those columns are missing (~98%) or the 20 testcases have all NA on those columns, so they can be removed to reduce computation difficulties

```
pml.training <- pml.training[c(1:4, 30:42, 53, 61, 77:79, 95, 106:117, 133, 144:153)]
pml.testing <- pml.testing[c(1:4, 30:42, 53, 61, 77:79, 95, 106:117, 133, 144:153)]
```

This reduces the number of columns from 153 columns to 46.

Setting up a training and testing set for model selection

To select a model, first the training dataset should be converted to a training part and validation part. To make the results comparable after serveral runs of the script a seed is set.

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.4.4
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
set.seed(5678)
Trainpart = createDataPartition(pml.training$classe, p = 0.3)[[1]]

training = pml.training[Trainpart,]
validation = pml.training[-Trainpart,]
```

Model creation

Two different prediction models are explored: Recursive Partitioning and Random Forest. After fitting a prediction model, the accurarcy will be validated on the validation dataset.

```
rpart_mod <- train(classe~., method="rpart", data=training)
```

```
rf_mod <- train(classe~., method="rf", data=training)
```

Now the models are created the validation of the models can be done by running the models on the validation dataset.

```r
predict_rpart <- predict(rpart_mod, validation)
confusionMatrix(predict_rpart, validation$classe)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 2378  414   66  111   39
##          B  421 1524  113  320  596
##          C  826  596 1883 1189  589
##          D  270  123  333  631  142
##          E   11    0    0    0 1158
##
## Overall Statistics
##
##                Accuracy : 0.5515
##                  95% CI : (0.5432, 0.5599)
##     No Information Rate : 0.2844
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4378
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.6088   0.5736   0.7862  0.28032  0.45880
## Specificity            0.9359   0.8691   0.7178  0.92440  0.99902
## Pos Pred Value         0.7906   0.5124   0.3705  0.42095  0.99059
## Neg Pred Value         0.8575   0.8947   0.9408  0.86758  0.89128
## Prevalence             0.2844   0.1935   0.1744  0.16391  0.18379
## Detection Rate         0.1732   0.1110   0.1371  0.04595  0.08432
## Detection Prevalence   0.2190   0.2166   0.3701  0.10915  0.08512
## Balanced Accuracy      0.7723   0.7213   0.7520  0.60236  0.72891
```

```r
predict_rf <- predict(rf_mod, validation)
confusionMatrix(predict_rf, validation$classe)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 3879   47    0    0    0
##          B   18 2580   43    1    5
##          C    5   28 2339   56   15
##          D    3    1   12 2194   18
##          E    1    1    1    0 2486
##
## Overall Statistics
##
##                Accuracy : 0.9814
##                  95% CI : (0.979, 0.9836)
```

```
##     No Information Rate : 0.2844
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9765
##  Mcnemar's Test P-Value : 2.646e-14
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.9931   0.9710   0.9766   0.9747   0.9849
## Specificity            0.9952   0.9940   0.9908   0.9970   0.9997
## Pos Pred Value         0.9880   0.9747   0.9574   0.9847   0.9988
## Neg Pred Value         0.9972   0.9931   0.9950   0.9950   0.9966
## Prevalence             0.2844   0.1935   0.1744   0.1639   0.1838
## Detection Rate         0.2825   0.1879   0.1703   0.1598   0.1810
## Detection Prevalence   0.2859   0.1927   0.1779   0.1622   0.1812
## Balanced Accuracy      0.9942   0.9825   0.9837   0.9859   0.9923
```

Evaluating the accuracy of all models the random forest model has the highest accuracy.

Conclusion

Since the accuracy of the random forest model is the highest, this model will be choosen to make the predicitons for the quiz related to this course.